Knowledge and Understanding of Asia: Using a Common Item Pool to Gain a National Picture

Patrick Griffin and Kerry Woods
Assessment Research Centre
Faculty of Education
The University of Melbourne

Paper presented at the Annual Conference of the Australian Association for Research in Education, Auckland, December 2003.

Abstract

A series of tests, attitude and questionnaire scales were developed to assess the proficiency of Australian Year 5 and Year 8 students in studies of Asia. Questionnaires were administered to students, teachers and school principals to identify appropriate issues related to outcomes in terms of classroom, curriculum, and teaching and learning practices. This paper presents results of analyses that involved calibrating items distributed over 14 overlapping subtests, developed to cater for state and territory curricula and two year levels. This allowed for state and year level preferences to be selected from a common pool of 105 items representing the three key learning areas of Studies of Society and Environment (SOSE/HSIE), English and the Arts. The project used common item anchoring to map all students and items onto a single, underpinning scale that was identified and interpreted using concurrent equating procedures and a skills audit of items.

In Australia, schools in each state and territory have been implementing studies of Asia programs since 1993. Research designed to identify the knowledge, understanding and attitudes of Australian school child ren in the area of studies of Asia was first mooted in 1999, and in 2001 the Federal Minister for Education asked for a study to establish a national database. The National Asian Languages and Studies in Australian Schools (NALSAS) Taskforce later took over the funding and governance of the study. A series of tests, attitude and questionnaire scales were developed to assess proficiency in studies of Asia as designated by the Asia Education Foundation (AEF). Questionnaires were administered to students, teachers and school principals to identify appropriate issues related to outcomes in terms of classroom, curriculum, and teaching and learning practices.

Objectives and Overview

A primary goal was to measure proficiency in studies of Asia at two formative stages of development, Years 5 and 8 of schooling. This aspect of the project involved development of tests consistent with a generic content of studies of Asia. In addition, the study sought to identify factors related to proficiency. These included the context in which students developed their knowledge and attitudes, encompassing the classroom, home and general educational milieu. The study was cross sectional in design, and so interpretations of a longitudinal nature were not proposed. Specifically the aims of the project were to

- collect and analyse national data on Year 5 and Year 8 students' knowledge about Asia and their attitudes to learning about Asia;
- provide states and territories with calibrated achievement tests of students' knowledge and understanding of Asia; and
- provide the means to extend beyond this project by identifying patterns of educational context that influence students' proficiency.

To achieve these aims, the project was required to develop a total of 14 overlapping tests to cater for state and territory curricula, and two year levels. The solution involved a simpler approach in that a single pool of items was developed and calibrated. This paper presents part of the analyses that involved calibrating items distributed over the 14 subtests, allowing for state and year level preferences to be selected from the common pool of 105 items representing the three key learning areas of Studies of Society and Environment (SOSE/HSIE), English and the Arts. Using concurrent equating procedures, a single dominant dimension was identified in this set of analyses.

The NALSAS Strategy.

The NALSAS Strategy, a cooperative initiative of the Commonwealth, state and territory governments, assisted schools to improve participation and proficiency in four targeted Asian languages: Japanese, Modern Standard Chinese, Indonesian and Korean. It also supported studies of Asia across the curriculum. The Strategy encouraged expanded provision of Asian languages and Asian studies through all school systems in order to improve Australia's capacity and preparedness to interact internationally, in particular with key Asian countries.

Commencing in schools in 1996, Strategy funds were initially applied to activities such as development and production of curriculum and teaching materials, resources, teacher training and professional development. The resources committed to the Strategy by Commonwealth and state and territory governments, have been substantial. From inception to the end of 2002, the Commonwealth provided over $208 million to support the Strategy. Commonwealth funding was paid to government education authorities, Catholic Education Commission Offices and Associations of Independent Schools on a per capita basis of school enrolments. The four agreed focus areas of the NALSAS Strategy were curriculum delivery, teacher quality and supply, strategic alliances, and outcomes and accountability.

Access Asia schools.

The Access Asia school program began with a total of 80 schools in 1993. Today the program has grown to approximately 1800 schools, supported by an injection of NALSAS funds from both national and states/territory levels. There is considerable variation between states and other jurisdictions regarding the meaning of involvement in the program in that each system of education interprets its own way of being involved.

The AEF encourages whole school planning, although this is not always possible given resourcing issues and levels of commitment to studies of Asia by schools. The AEF's evaluators have found that the longer a school participates in the Access Asia program, the more likely it is that the school will meet expectations regarding whole school change. However, expectations of school commitment have not been documented, nor have there been any previous attempts to identify student outcomes associated with studies of Asia programs and their various interpretations at school and state level. This was the first such attempt, but this project was restricted to a national, rather than state, level examination of outcomes and related predictors. It should be noted, therefore, that variation between states in interpretation of the Access Asia program was not investigated. However, data collected from each state were appropriately weighted to take into account relative contributions to national norms.

Method

Sampling

The target population was defined as all schools in all states and territories. A four-stage cluster sample was selected (Asia program within state, school within program (within grade level) and class within school) with schools chosen using probability proportional to size. A student cluster size of 25 was used for sampling purposes and for calculating sampling errors. Extrapolating the 1997 figures supplied by Baumgart and Halse (1999) provided estimates of numbers of Access Asia schools in each state. A similar sampling procedure was adopted for the non Access Asia schools. Details of the sample are provided in Table 1.

Table 1
The Sample

	Primary				Secondary				Total	
	Access Asia		Non Access Asia		Access Asia		Non Access Asia			
	School	Student	School	Student	School	Student	School	Student	School	Student
NSW	13	306	14	384	13	323	17	429	57	1442
VIC	20	441	17	385	21	477	29	708	87	2011
QLD	4	101	12	277	1	29	19	464	36	871
SA	8	248	11	282	11	274	3	38	33	842
WA	15	375	7	196	18	425	10	345	50	1341
TAS	7	189	0	0	5	111	2	25	14	325
ACT	5	131	0	0	1	16	1	63	7	210
NT	3	65	3	86	7	146	0	0	13	297
Total	72	1856	64	1610	77	1801	81	2072	297	7339

Students from two separate year levels (5 and 8) participated in the research to highlight emerging and consolidating attitudes. Year 8 students were surveyed in order to include a secondary school sample and to accommodate the participation of secondary classes in Western Australia and Queensland. The selected year levels provided information about development at critical stages in primary and secondary education. They also enabled a cross sectional comparison of attitudes and knowledge between students at two stages of schooling.

Instrumentation

The project sought to identify levels of proficiency based on direct interpretation of the cognitive skills underpinning responses to test and questionnaire items. Identified skills were converted into descriptors and used to ascertain distributions for national norms. The measurement of proficiency was guided by a curriculum framework supplied in support documents produced by the AEF that included pointers, examples of curriculum sequences and student work samples in key learning areas. These were matched to curriculum frameworks and outcome statements provided by state and territory education systems. Data on related context issues were collected using principal, teacher and student questionnaires. Questionnaire content was determined through consultation with representatives of the AEF, Commonwealth and state and territory education systems. A survey of representatives was conducted to identify policy-related questions and to finalise the format and content of each questionnaire. In addition, a major source of issues for questionnaire items was the Baumgart and Halse (1999) report.

Test development.

Specifications of instruments were set and agreed upon by all state and territory educational jurisdictions. An analysis of curriculum materials in the national profiles, state curriculum guidelines and syllabus statements, and the NALSAS and AEF curriculum materials, guided areas to be included in tests. These investigations directed development of specifications and selection of learning outcomes for assessment. The mapping exercises identified those elements of studies of Asia curriculum outcomes that were considered nationally important, and these helped to form blueprints for the test design and determined selection of curriculum and prompt materials. All states and territories agreed that each test should have about 60 questions and that an appropriate amount of time should be made available in schools for testing. Items were

written to reflect the knowledge and understanding of Asia that could be expected of students in Years 5 and 8.

A reference group of specialists for studies of Asia in each state and territory was established. Workshops were conducted to ensure a sample of items was developed that matched local state and territory curricula. Specific subject specialists were engaged to write test items, and these personnel were drawn from teachers and consultants in each state and territory. A specific decision was taken to use multiple choice items in the investigation of students' knowledge. Although multiple choice items had not been used widely in studies of Asia there were various reasons for the decision, including observations that multiple choice items allowed objectivity of marking and had been used with success in other state assessments.

Each test item had four response options, consisting of one right and three wrong answers. The disadvantage of this format was that students had a 25% chance of guessing correctly. The advantage, however, was that only three good wrong answers need be created. This was an instruction to item writers. All wrong answers were expected to provide information to curriculum developers. A set of 15 items was specified as an overlap between the Years 5 and 8 tests, to enable tests to be linked and to help describe differences in achievement.

Table 2
The Proposed Construct and Framework for Item Development

	English	Art	SOSE/HSIE
D	In relation to Asia, complex ideas, information and advertisements; conjecture and prediction based on extension beyond given text. Infer writer's purpose and values. Compare text and styles in an Asian context.	Links between music, art and culture. Reasons for specific presentation of art and change of art. Influences of external pressures on Asian art forms.	Able to deal with and identify different cultural elements in complex settings and to show the effects of cultural shifts,
C	Locate, translate and/or interpret information about Asia. Cross reference ideas and information about Asia; compare cultural ideas and information and link them.	Understands the influence of religion and politics on the presentation of art and how individual people respond to art forms.	Links the culture of the country to different patterns of behaviour, religions and customs.
B	Locate single idea information in the text using specific Asia-related prompts based on common knowledge and stereotypes with specific words, phrases, sentences and icons in text.	Classifies artistic representations by region and county. Can recognise ways that common articles and people are represented through common art forms	Understands effect of weather, national dress and some common customs in selected countries.
A	Recognise single item of information or icon; identify information in word arrangements, and use of simple cues such as vocabulary and icons.	Recognition of simple art forms, using simple lists and commonly known examples	Names locations and countries, places on maps, general simple knowledge

Table 2 shows the proposed progression within each of three key learning areas presented to participants at item writing workshops. Item writers were encouraged to link curriculum outcomes to show how understanding progressed, via the content provided in state level curriculum documents. The framework shows four broad levels of development, labelled from Levels A to D.

Data collection was conducted in one school year and the project used common item anchoring to map all students and all items onto a single underpinning scale. In addition, each state and territory identified a subset of items from each test that was most related to that state or territory's curriculum, and student ability was only estimated using these items. That is, all students in a Year level took the same test but each student's performance was determined using a subset of items confirmed as relevant by state and territory curriculum specialists.

Instrument Development

The following sections detail the various scale and test items that were estimated using item response modelling. More specifically, the Quest computer program (Adams & Khoo, 1995) was used to apply a simple Rasch model to items used in the Years 5 and 8 tests of knowledge and understanding.

The term *ability* often leads to emotive and confused discussions of exactly what is being measured, and the dimensions and fairness of such a term. In this case, ability is defined as the level of development within a domain being measured. It is not a fixed property and can be altered by targeted intervention.

Test data were analysed using both classical and item response (Rasch) analyses. Item response modelling was used to equate across systems and identify student ability estimates. By comparison, raw scores would be misleading as each state and territory selected their own subset of items. Item response modelling enabled abilities of students from each jurisdiction to be mapped onto a common scale, even though student groups were being assessed with different subsets of items. The common scale made it possible to analyse student performance in terms of other background characteristics obtained through student, teacher and principal survey instruments. These analyses have been reported in detail by Griffin and others (2002).

Student Measures

In two tests of 60 items, each covering several levels of ability across three learning areas, it was not possible to have a lot of items in every cell of the test blueprint. Measures of domains of each curriculum area could have been derived from test scores, but their (alpha) reliability would be expected to be low and errors of measurement would be high. Analyses of student performance, therefore, focus on one main measure - the knowledge performance measures. However, two measures are used in this report to describe student performance. The first is a transformed score in which ability estimates were standardised with a mean of 500 and a standard deviation of 100 (K500 score). The second is competence level identified through a skills audit of items and an interpretation of the groups of items on the underpinning scale.

Scaling

Test items were scored as either right or wrong. Scoring each item in this manner treats them as independent dichotomous items, in which each student, n, has an ability $?_n$ and each item has a difficulty parameters ?? representing the difficulty of attaining a score of 1 on each item 1 to ?. Each of these parameters governs the likelihood of a student with ability, $?_n$, obtaining a score of 1 rather than 0. The analysis models the relationship between student ability and the difficulty parameters of each of the items.

The Rasch simple logistic model, using the computer program Quest (Adams & Khoo, 1995) and the RUMM2010 package (RUMM Laboratory, 2001), was used to derive estimates of item difficulty and student knowledge. The probability of a correct response was obtained by

$$\Pr\{x ? 1 \backslash ?_n, ?_i\} ? \frac{e^{(?_n ?? _i)}}{1 ? e^{(?_n ?? _i)}}$$

The probability level for comparisons was set at 0.5 because of the interpretation and instructional implications of this value.

The two packages were used because they are based on different estimation procedures. Quest uses a marginal maximum likelihood approach and RUMM uses a pairwise procedure based on Zwinderman's algorithms (1995). The probabilities (that the score was x=1 for an item *i*) enabled estimates of ability $?_n$ and difficulty parameters $?_i$ to be obtained. These estimates were simultaneously plotted on a variable map illustrating the relative position of students against difficulty levels assigned to each test item.

After measurement properties of items had been established, subsets of items were used to measure the latent variable. It was not necessary to use all items during any particular assessment to obtain a measurement of student ability. Items were selected to target curriculum emphases for specific educational jurisdictions, as shown in Table 3. Despite differences in emphases for year levels or systems, selected items were proposed to measure the same underlying variable, yet provide specific information for different curricula.

Table 3

Item Selection for each State and Territory

	Primary Test Item Selection								Secondary Test Item Selection							
	VIC	NSW	ACT	QLD	TAS	SA	WA	NT	VIC	NSW	ACT	QLD	TAS	SA	WA	NT
1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2		1	1	1	1	1		1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1
4			1	1				1		1	1	1		1	1	1
5	1	1	1	1	1	1	1	1			1					1
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1
7	1	1	1		1	1	1	1	1		1					1
8		1	1	1		1		1	1		1	1		1	1	1
9	1	1	1	1	1	1	1	1	1	1	1			1	1	1
10	1	1	1	1	1	1		1			1	1		1		
11		1	1	1	1	1	1	1	1	1	1		1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1		1	1	1	1		1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1		1		1	1	1			1	1
15	1	1	1	1		1	1	1	1	1	1			1		1
16	1	1	1	1	1	1		1			1	1		1		1
17	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1
18	1	1	1	1	1	1	1	1	1	1	1			1	1	1
19			1	1		1		1	1		1	1	1	1	1	1
20	1	1	1	1	1	1		1	1	1	1		1	1	1	1
21		1	1		1	1	1	1	1		1			1	1	1
22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23		1	1	1		1	1	1		1	1	1	1	1	1	1
24			1					1	1	1	1	1	1	1	1	1
25	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1
26			1			1		1	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1
29		1				1	1	1	1	1	1	1	1	1	1	1
30		1	1	1		1		1	1	1	1	1	1	1	1	1
31	1			1		1		1		1	1	1	1	1	1	1
32			1	1	1	1		1		1	1	1		1	1	1
33	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1
34	1	1	1		1	1	1	1			1			1	1	1
35	1	1	1		1	1	1	1			1			1	1	1
36	1	1	1		1	1	1	1		1	1			1	1	1
37	1	1	1	1		1	1	1		1	1		1	1	1	1
38	1	1	1		1	1		1		1	1	1		1		1
39	1	1	1			1		1	1	1	1	1	1	1	1	1
40	1		1	1	1	1	1	1	1		1	1		1	1	1
41	1		1	1		1		1	1	1	1		1	1	1	1
42	1		1	1		1		1		1	1	1	1	1	1	1
43	1			1			1	1	1	1	1	1	1	1	1	1
44		1	1			1		1	1	1	1	1	1	1		1
45		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
46	1	1	1	1	1	1	1	1	1	1	1			1	1	1
47	1	1	1	1	1	1	1	1	1		1		1	1	1	1
48	1	1	1	1	1	1	1	1		1	1	1	1	1		1
49	1	1	1	1	1	1	1	1	1	1	1	1		1		1
50	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1
54	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1
52		1	1	1	1	1	1	1	1		1	1		1	1	
53		1	1	1	1	1			1		1	1			1	1
54		1	1			1		1						1		1
55	1		1	1	1	1	1	1	1		1			1	1	1
56	1	1			1	1		1		1	1	1			1	1
57	1	1	1	1	1		1	1		1	1	1		1	1	1
58	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1
59	1	1	1	1		1		1	1		1	1		1	1	1
60	1	1	1	1	1	1	1	1	1		1			1	1	1

After all item selections were taken into account, the test items were calibrated using concurrent equating. While there are some disadvantages to this procedure, its simplicity and ease of application saves a great deal of complex data file manipulation. One of the problems with this method is the amount of missing data. When items are rejected at a system level, they are treated as though they have not been administered. Students also omit items, and hence add to the proportion of missing data. Previous analyses of missing data in similar data sets such as the Southern African Consortium for Monitoring Educational Quality (SACMEQ) study (Coates, 2001; Griffin, 1999) suggested that when the test is timed appropriately, decisions made by the student to leave items out and not respond should be treated as incorrect responses. Differences in ability estimation become important, under these guidelines, when the proportion of missing data exceeds 8%. No item in this test had such a large proportion of missing data as a result of student decision, although there were many items with larger proportions of missing data because of systemic decisions not to select the item. The result of systemic decisions not to administer the item (or to have it treated as non-administered) was that 14 data sets were merged. When these were conjointly calibrated the result was a series of discrete subtest characteristic curves showing the relationships between the raw scores and logits. Figure 1 illustrates these relationships.

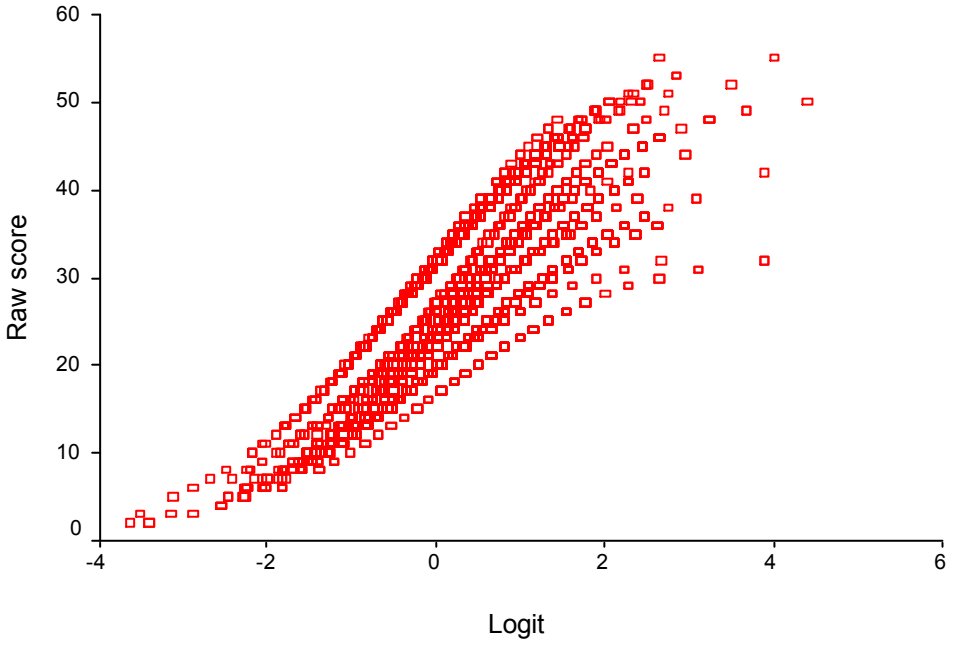

Figure 1. Relationships between raw scores and logits for 14 subtests.

Despite the complexity of relationships, it can be seen that there is a clear indication that raw scores have different interpretations in terms of student ability. For example, a raw score of 30 translates to a logit range of –0.2 to +2.5. The difference in terms of knowledge and understanding of Asia would thus be substantial, illustrating both the need for, and power of, equating the subtests.

Several measures of accuracy of the test data were used. The first was the measure of the standard error of measurement for each of the item difficulty estimates. A second was a measure

of the extent to which the data fit the Rasch (1980) model. This measure is the mean squared differences between the estimated or modelled difficulty and the observed difficulty of each score point, weighted by the variance of the assigned scores. This is called the INFIT mean square and this stands for the Information Weighted Mean Squared deviation score. The expected value of the INFIT is 1.0 (Adams & Khoo, 1995).

Fit to the model helped identify how well the item response patterns and student response patterns conform to the assumptions underlying the model. Items may misfit if students' response patterns to those items are related to extraneous factors, rather than to ability on the trait being measured. Student responses may misfit if students respond incorrectly to items below their ability or give an unexpected correct answer to a very difficult item that is above their ability. There are several explanations, such as guessing, for this type of test behaviour.

Table 4 presents the summary calibration statistics for the item pool. Some things are particularly noteworthy. The mean item difficulty is arbitrarily set to zero. In this case, variance of item difficulty levels was 1.00 with a reliability of item separation of 1.00. The mean item INFIT was 0.99 with a standard deviation of 0.07. There were no items with zero scores and no items with perfect scores. Mean student ability estimate was 0.24, indicating student ability level was slightly higher overall than test difficulty. Standard deviation of student ability estimates was 0.87, slightly less than variance of item difficulties, and this indicated that the task was well matched to the range of student abilities, particularly at the upper end of the ability range. The reliability of the student separation index was 0.89. Mean squared INFIT index was 1.00 with a variance of 0.13, indicating the test was well matched to the majority of students and that a single dominant latent variable underpinned the set of items. It also suggests the test successfully separated students on the basis of ability (i.e. it possessed acceptable criterion validity) as well as demonstrating construct validity. However, there was no external evidence available regarding the nature of the construct criterion.

Table 4

Calibration Summary for the Item Pool

	Mean	Std. dev
Item Separation	1.00	
Item INFIT	.99	.07
Item Difficulty	.00	1.00
Student Ability	.24	.87
Student Separation	.89	
Student INFIT	1.00	.13

Tables 5 and 6 present the calibration characteristics for the 105 items used in the tests. Descriptive data for each item include item name, the number of tests the item is in, the number of Year 5 and Year 8 students responding to the item, and the proportion of students responding to each alternative. The statistical analyses reported include the difficulty level (logit) and the fit to the item response model used to calibrate the test (expected value between 0.77 and 1.3). It should be noted that INFIT values were all within the range of 0.77 to 1.3, thus providing additional evidence of a dominant underlying dimension in the variable being measured.

Table 5
Primary Test Item Calibration

Item	KLA	Item x Test	Cases	Student Response (%)				Logit	Infit Mean Square	SEM
				A	B	C	D			
P1	S	1	2610	2.30	2.90	5.10	89.30	-2.49	1.04	0.04
P2	S	1	2039	1.80	92.00	2.60	3.20	-3.14	1.19	0.05
P3	S	1	3414	6.90	6.00	83.00	3.60	-1.88	.99	0.03
P4	E	1	2124	22.80	22.50	20.30	33.00	.82	1.11	0.03
P5S1	S	2	7287	8.60	4.30	81.50	5.00	-1.53	1.04	0.03
P6S2	S	2	7287	1.80	2.90	1.90	92.80	-3.14	1.15	0.04
P7S11	S	2	6416	9.40	5.70	74.50	9.60	-1.10	.94	0.04
P8	S	1	1850	7.00	82.10	4.60	5.80	-1.75	.95	0.05
P9	S	1	3414	79.10	14.00	2.80	3.60	-1.60	.95	0.04
P10	S	1	2843	19.90	70.70	5.30	3.70	-1.06	1.03	0.04
P11	S	1	2610	11.60	43.40	24.50	19.70	.28	.98	0.03
P12	S	1	3414	43.00	7.80	6.40	42.20	.32	.97	0.04
P13	S	1	2843	45.00	13.20	17.90	23.00	-.04	.99	0.04
P14	S	1	2843	13.20	59.80	14.70	10.90	-.49	.97	0.04
P15	E	1	3225	51.80	18.60	13.10	15.90	-.09	.96	0.03
P16	S	1	2843	16.40	21.80	52.00	9.40	-.13	.97	0.04
P17	S	1	3414	18.20	58.60	14.50	8.30	-.43	.99	0.04
P18	S	1	3414	49.40	10.60	18.00	21.50	.01	.97	0.04
P19	E	1	1850	7.90	34.50	10.80	46.10	.50	1.02	0.04
P20S6	A	2	5946	9.40	1.80	1.40	86.90	-2.01	.98	0.03
P21	E	1	2232	3.50	87.70	2.70	5.60	-2.33	1.01	0.04
P22S33	S	2	7287	8.90	9.00	9.00	72.10	-.86	.91	0.04
P23S4	S	2	4973	9.70	17.30	58.20	14.10	-.08	1.09	0.05
P24S5	S	2	1919	13.90	3.70	75.00	6.70	-.92	1.03	0.05
P25	A	1	3225	55.80	14.50	14.50	14.40	-.27	1.11	0.03
P26	S	1	1472	4.00	4.20	31.70	59.40	-.44	1.05	0.03
P27S3	A	2	5946	44.20	32.40	8.60	13.50	.50	1.13	0.03
P28	S	1	2843	10.30	62.20	17.90	8.90	-.75	1.02	0.03
P29	S	1	1912	45.60	17.90	21.80	14.10	.12	.99	0.04
P30	E	1	1850	6.50	16.20	69.80	7.10	-1.06	.92	0.05
P31	S	1	1863	22.20	18.80	15.30	43.10	.30	1.07	0.04
P32	S	1	2039	17.30	28.80	13.00	39.60	.47	1.02	0.04
P33	S	1	2843	62.20	18.60	10.20	8.10	-.56	.96	0.03
P34	S	1	3036	16.40	54.20	18.20	10.20	-.23	.95	0.04
P35	E	1	3036	21.70	35.10	24.90	16.90	.73	.98	0.04
P36S20	E	2	6416	11.80	7.90	41.70	37.70	.62	1.01	0.04
P37	S	1	3225	20.10	16.50	49.10	13.30	.03	.98	0.04
P38	S	1	2465	20.90	42.60	28.30	7.50	1.00	1.04	0.04
P39S15	S	2	4750	42.60	18.10	13.30	24.60	.59	1.01	0.04
P40	S	1	2754	23.10	37.20	19.90	18.50	.57	1.13	0.04
P41	E	1	1994	22.20	22.70	42.80	11.40	1.44	1.06	0.04
P42	S	1	1994	20.40	16.90	32.80	28.50	.76	1.05	0.03

Item	KLA	Item x Test	Cases	Student Response (%)				Logit	Infit Mean Square	SEM
P43	E	1	2564	33.70	21.90	26.10	16.70	.74	1.02	0.04
P44	E	1	1472	33.00	18.80	35.90	10.80	.95	1.17	0.04
P45S42	E	2	5298	25.40	43.80	16.00	12.90	.51	1.07	0.04
P46S25	S	2	7287	62.40	9.30	4.90	22.40	-.29	.93	0.04
P47S26	S	2	7287	10.00	66.40	12.70	9.50	-.55	.96	0.04
P48S27	S	2	7287	46.30	12.20	23.20	16.50	.41	1.00	0.04
P49S28	S	2	7287	10.50	11.30	17.80	58.90	-.17	1.07	0.04
P50	S	1	3036	8.40	76.50	7.90	5.50	-1.36	.94	0.04
P51	S	1	3414	77.20	6.80	7.90	6.70	-1.41	.92	0.04
P52	A	1	2610	8.70	69.90	7.40	12.10	-.94	1.01	0.04
P53	E	1	1888	3.00	11.80	5.50	77.80	-1.46	.92	0.04
P54	E	1	1472	10.90	6.90	62.20	17.50	-.62	.92	0.04
P55	E	1	3414	36.80	18.00	26.40	15.80	.56	1.00	0.04
P56	S	1	2334	20.50	22.50	20.10	34.20	1.38	1.15	0.04
P57	S	1	2224	16.80	13.80	37.00	29.70	.58	1.03	0.04
P58	E	1	2754	19.00	34.30	31.30	12.70	.79	1.11	0.04
P59	E	1	2654	11.10	44.60	9.30	32.60	.17	1.00	0.04
P60	S	1	3414	6.00	59.50	14.90	17.50	-.43	1.01	0.05

Note. Items are coded according to their inclusion in the primary or secondary tests. Key learning areas include Studies of Society and Environment (S), English (E) and the Arts (A).

Table 6

Secondary Test Item Calibration

Item	KLA	Item x Test	Cases	Student Response (%)				Logit	Infit Mean Square	SEM
				A	B	C	D			
S1P5	S	2	7287	8.60	4.30	81.50	5.00	-1.53	1.00	0.04
S2P6	S	2	7287	1.80	2.90	1.90	92.80	-3.14	1.17	0.04
S3P27	A	2	5946	44.20	32.40	8.60	13.50	.50	1.12	0.04
S4P23	S	2	4973	9.70	17.30	58.20	14.10	-.08	1.08	0.04
S5P24	S	2	1919	13.90	3.70	75.00	6.70	-.92	1.01	0.04
S6P20	A	2	5946	9.40	1.80	1.40	86.90	-2.01	1.00	0.04
S7	S	1	2162	9.20	60.00	16.80	13.20	.05	1.10	0.04
S8	A	1	2985	13.80	5.50	78.20	1.80	-1.06	1.03	0.04
S9	E	1	3244	9.70	10.30	62.70	16.10	-.02	1.05	0.04
S10	E	1	1636	36.60	24.20	12.30	25.50	1.27	1.18	0.04
S11P7	S	2	6416	9.40	5.70	74.50	9.60	-1.10	.95	0.04
S12	S	1	3873	12.70	60.80	7.80	17.20	.00	.97	0.04
S13	S	1	3873	15.90	70.30	8.00	4.10	-.56	.97	0.04
S14	E	1	2240	15.80	16.10	52.00	14.50	.49	1.08	0.04
S15P39	S	2	4750	42.60	18.10	13.30	24.60	.59	1.00	0.05
S16	E	1	1782	7.70	66.80	16.10	8.50	-.30	.98	0.06
S17	S	1	3380	5.90	14.20	13.20	66.00	-.29	.95	0.07
S18	S	1	2492	5.20	72.90	7.30	13.70	-.75	1.01	0.05
S19	S	1	3873	6.50	25.00	9.60	57.50	.14	1.05	0.04
S20P36	E	2	6416	11.80	7.90	41.70	37.70	.62	1.00	0.05

Item	KLA	Item x Test	Cases	Student Response (%)				Logit	Infit Mean Square	SEM
S21	E	1	3244	49.10	17.40	24.10	7.80	.58	.99	0.05
S22	S	1	3873	5.40	7.30	70.80	15.60	- .58	.91	0.04
S23	S	1	2688	15.90	10.90	6.80	65.20	2.49	1.09	0.04
S24	S	1	3873	3.90	5.80	76.10	13.30	- .78	.91	0.04
S25P46	S	2	7287	62.40	9.30	4.90	22.40	- .29	.93	0.04
S26P47	S	2	7287	10.00	66.40	12.70	9.50	- .55	.96	0.04
S27P48	S	2	7287	46.30	12.20	23.20	16.50	.41	1.00	0.04
S28P49	S	2	7287	10.50	11.30	17.80	58.90	- .17	1.05	0.04
S29	A	1	3873	9.20	72.10	6.10	11.70	- .69	.96	0.04
S30	A	1	3873	3.10	83.00	6.00	7.00	-1.41	.93	0.04
S31	A	1	2688	12.20	10.80	25.40	50.80	.56	1.00	0.04
S32	A	1	2552	3.80	65.50	24.90	4.70	- .19	.99	0.06
S33P22	S	2	7287	8.90	9.00	9.00	72.10	- .86	.91	0.04
S34	E	1	1307	18.50	16.90	55.80	7.30	.20	.96	0.04
S35	E	1	1307	8.30	29.60	47.80	12.80	.63	1.08	0.04
S36	E	1	2059	61.20	12.60	16.00	8.30	- .09	.86	0.04
S37	S	1	2195	7.50	49.60	32.80	8.10	.66	1.13	0.04
S38	E	1	1782	5.90	62.20	15.90	14.10	.03	.99	0.04
S39	S	1	3873	22.90	9.50	30.00	35.80	1.21	1.05	0.04
S40	E	1	3737	34.70	27.10	15.70	20.10	1.29	1.06	0.04
S41	S	1	3380	39.60	29.80	19.60	8.70	1.00	1.01	0.04
S42P45	E	2	5298	25.40	43.80	16.00	12.90	.51	1.04	0.04
S43	S	1	3873	26.90	14.00	40.50	15.70	1.02	1.05	0.04
S44	E	1	3103	10.30	24.20	19.50	43.70	.84	.93	0.04
S45	S	1	3873	46.10	12.80	25.30	13.50	1.86	1.15	0.04
S46	A	1	3244	26.60	15.00	25.70	30.30	1.50	1.13	0.05
S47	E	1	3380	21.60	16.20	12.90	46.70	.66	.91	0.04
S48	A	1	3103	7.70	9.70	72.50	7.20	- .56	.87	0.04
S49	A	1	2967	34.60	11.10	12.20	38.90	1.32	1.07	0.04
S50	S	1	3873	13.30	15.70	8.40	59.20	.02	.94	0.05
S51	S	1	3380	23.70	43.00	18.40	11.10	.94	1.06	0.05
S52	A	1	2839	35.70	20.40	14.60	25.10	1.62	1.03	0.04
S53	A	1	2673	9.00	72.60	7.50	7.10	- .63	.95	0.05
S54	S	1	458	19.60	30.00	18.80	27.10	2.13	1.33	0.04
S55	S	1	2492	18.70	18.40	25.10	33.50	1.84	1.17	0.04
S56	E	1	2240	62.60	11.40	12.70	8.80	- .23	.90	0.05
S57	E	1	2552	23.70	32.70	32.40	6.50	1.45	1.02	0.04
S58	E	1	2688	66.60	10.60	9.70	8.50	- .43	.85	0.04
S59	E	1	3737	23.00	40.40	11.00	21.50	1.04	1.09	0.04
S60	A	1	2492	46.90	19.20	14.30	15.30	2.20	1.06	0.04

Note. Items are coded according to their inclusion in the primary or secondary tests. Key learning areas include Studies of Society and Environment (S), English (E) and the Arts (A).

All test items were mapped onto a common scale and ordered according to difficulty. Results of the analysis are shown in Figure 2. On the left of the figure there is a scale ranging from –3.0 to +3.0. This is the logit scale or measure of student ability and item difficulty. The next part of the figure is a distribution of students with the symbol 'X' used to represent approximately 30 students at each level. The chart presents items in order of increasing difficulty. As indicated by the item difficulty distribution, the secondary test was slightly harder than the primary test and the link items were spread over the lower range of difficulty.

Logit	Students	Secondary	Primary	Link
	X	s23		
	X			
	X	s60		
2.0	X			
	XX			
	XXX	s45 s52 s55		
	XXXXXXX			
	XXXXX	s46 s54	p41 p56	
	XXXXXXXXX	s40 s49 s57		
	XXXXXXXX	s10 s39 p38	p38	
	XXXXXXXXXX	s41 s43 s59		
1.0	X			
	XXXXXXXXXX	s44 s51	p4 p42 p43 p44	
	XXXXXX			
	XXXXXXXX	s47	p19 p35 p55 p57	s20p36
			p58	
	XXXXXXXXXX	s21 s35 s37	p32 p40	s3p27 s15p39
	XXXXXXX			s42p45
	XXXXXXXXXX	s14 s31	p12 p31	s27p48
	XXXXXXX			
	XXXXXXXXXX	s34	p11 p13 p29 p59	
	XXXXXXX			
	XXXXXXXXXX	s7 s19 s50	p18 p37	
0.0				
	XXXXXXXXXX	s9 s12 s36 s38	p15 p16	s4p23
	XX	s56		
	XXXXXXXXXX	s32	p25 p34	s28p49
	XXXXX			
	XXXXXXXXXX	s16 s17 s58	p17 p26 p60	
	XX			
	XXXXXXX	s13 s22	p14 p28 p33 p54	s25p46 s26p47
	XXXXXXXXX	s18 s29 s48 s53		
		s24		s33p22
	XXXXXXXXX			
		s8	p10 p30 p52	s5p24 s11p7
-1.0	XXXXXX			
	XXXX			
	XXXX		p50	
	XX	s30	p51 p53	s1p5
	XXX		p9	
	X		p8	
	X		p3	s6p20
-2.0				
			p21	
			p1	
				s2p6
			p2	
-3.0				

Figure 2. Variable map for the knowledge and understanding tests.

The three sub-domains of the test were also well represented, although the Arts had fewer items and SOSE/HSIE tended to dominate test composition. Nevertheless, as shown in Figure 3, each domain covered a range of difficulty suitable for the student ability range and clustered to aid in interpretation.

Logit	Students	English	SOSE/HSIE	Arts
	X		s23	
	X			
	X			s60
2.0	X			
	XX			
	XXX		s45 s55	s52
	XXXXXXX			
	XXXXX	p41	s54 p56	s46
	XXXXXXXXX	s40 s57		s49
	XXXXXXXXX	s10	s39 p38	
	XXXXXXXXXXX	s59	s41 s43	
1.0	X			
	XXXXXXXXXXX	s44 p4 p43 p44	s51 p42	
	XXXXXX			
	XXXXXXXXX	s20p36 s47 p19 p35 p55 P58	p57	
	XXXXXXXXXXX	s21 s35 s42p45	s15p39 s37 p32 p40	s3p27
	XXXXXXXX			
	XXXXXXXXXXX	s14	s27p48 p12 p31	s31
	XXXXXXXX			
	XXXXXXXXXXX	s34 p59	p11 p13 p29	
	XXXXXXXX			
	XXXXXXXXXXX		s7 s19 s50 p18 p37	
0.0				
	XXXXXXXXXXX	s9 s36 s38 s56 p15	s4p23 s12 s25p46 p16	
	XX			
	XXXXXXXXXXX		s28p49 p34	s32 p25
	XXXXXX			
	XXXXXXXXXXX	s16 s58	s17 p26 p60 p17	
	XX			
	XXXXXXXX	p54	s13 s22 s26p47 p28 p14 p33	
	XXXXXXXXX		s18 s24 s33p22	s29 s48 s53
	XXXXXXXXX			
		p30	s5p24 s11p7 p10	s8 p52
-1.0	XXXXXXX			
	XXXX			
	XXXX		p50	s30
	XX	p53	s1p5 p51	
	XXX		p9	
	X		p8	
	X		p3	s6p20
-2.0				
		p21		
			p1	
			s2p6	
		p2		
-3.0				

Figure 3. Variable map showing distributions of students and items for key learning areas.

The two approaches to calibration and fit were used for specific reasons. During analyses of other data in international studies, such as SACMEQ (Griffin, 1999; Ross 1995), systematic differences were identified in the ability and difficulty estimates when concurrent equating

procedures used either Quest or Conquest, (based on marginal maximum likelihood procedures) or RUMM (based on pairwise procedures). Andrich and Lau (2002) identified that the removal of misfitting persons (and items) from the data for purposes of calibration provided a stable data set and convergence of ability and difficulty estimates from the two procedures. The two estimation procedures were thus used in this study to determine whether the same issue was present. It was not. Figure 4 shows the correspondence between the marginal and pairwise estimation procedures.

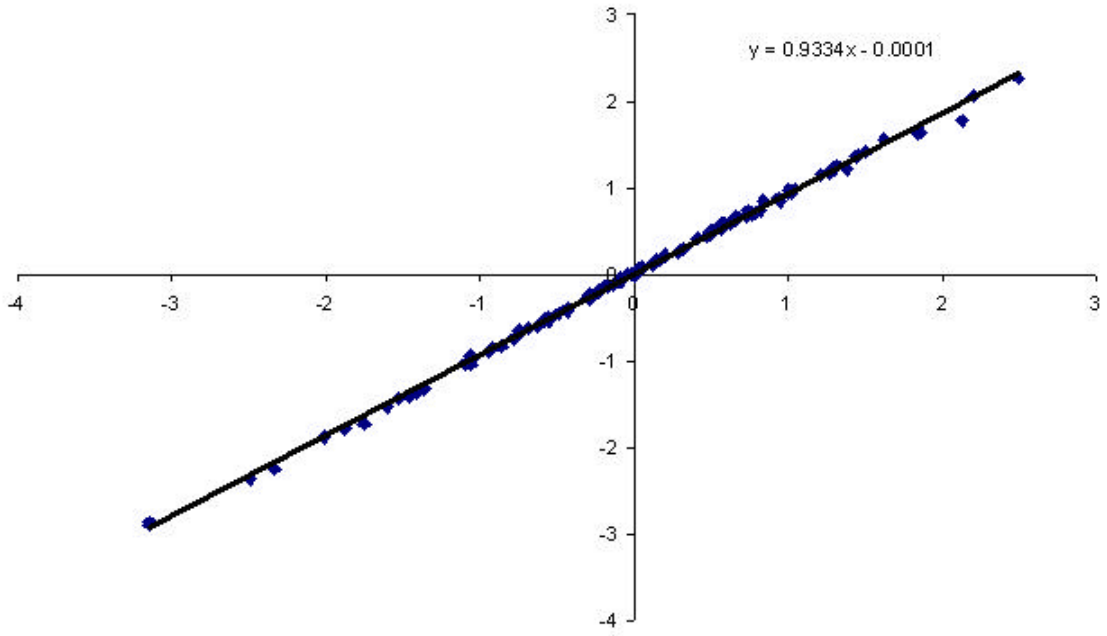

Figure 4. Comparison of item difficulty estimates using RUMM and Quest.

Despite the consistency of the two approaches, misfitting student data sets were removed from the data and the tests recalibrated. An item anchor set was then used to rerun the analysis using Quest including all students and items. Only one item misfitted in the calibration run and approximately 3% of students were found to have misfitting response patterns. These analyses also proved valuable for examination of the misfit.

Figures 5 and 6 show proportions of misfitting students in terms of a range of contextual factors, such as location of school, gender, year level, and ethnic background.

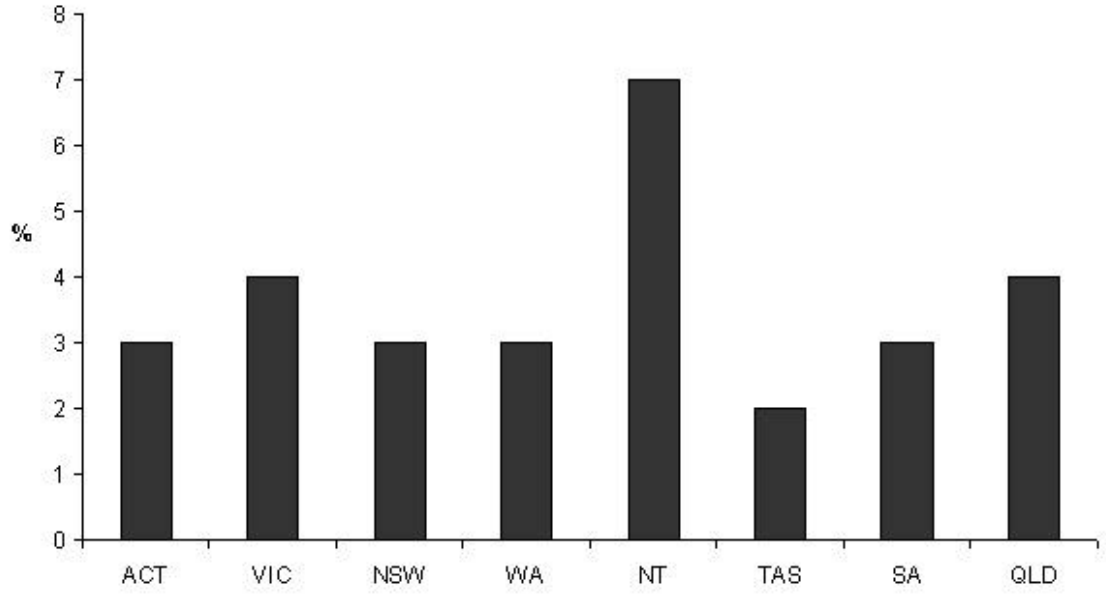

Figure 5. Percent of misfitting student response sets by state.

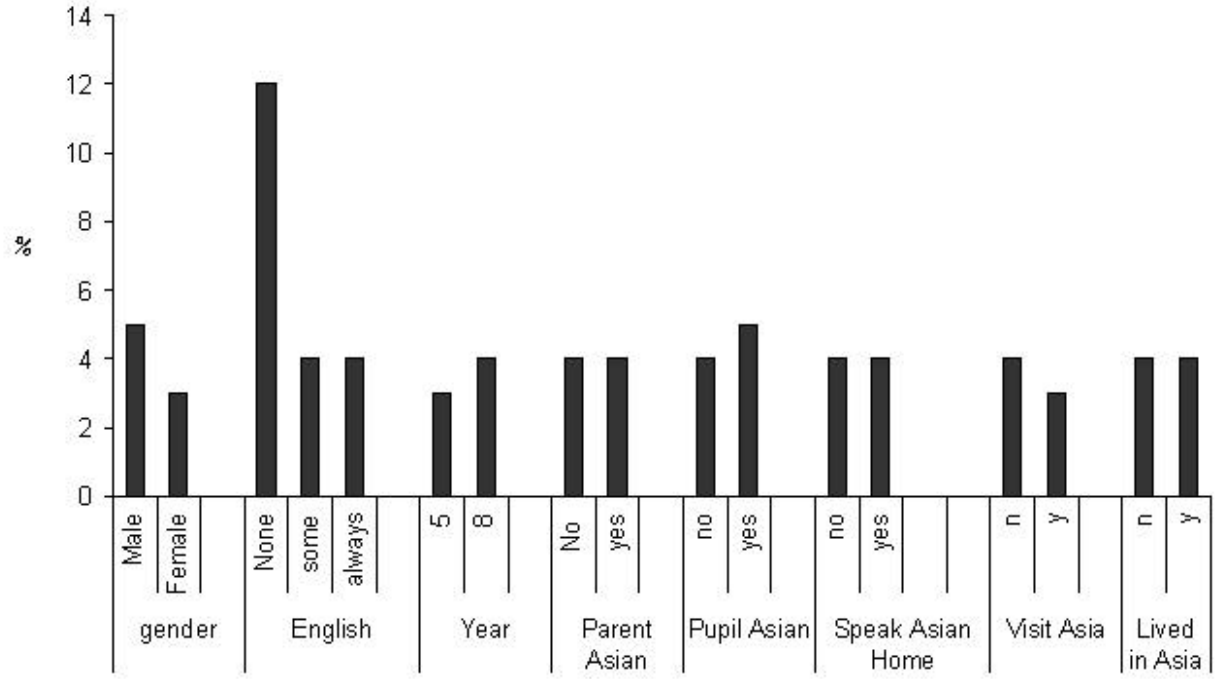

Figure 6. Percent of misfitting students by contextual variables.

It can be seen that the student misfit is generally unrelated to background factors. The exceptions are speaking English at home and school location in the Northern Territory. The latter still has better than 93% of student response patterns fitting the model and is not of concern, but the 12% of non English speaking students for whom the test was unsuitable perhaps ought to be removed from further analyses of the data

Validity was examined using two approaches. The first involved an analysis of skills underpinning the development of ability within the domain as demonstrated by students. The second used the approach outlined by Wright and Masters (1983), based on item and person separation indices and their reliabilities. The variable maps suggest that items tended to group together at different points along the scale. The underlying skills associated with these item clusters were examined to determine whether they could be interpreted as having a common level of development. This requires a form of "artistic" interpretation together with a down-to-earth knowledge of "how the students think" but permits us to understand the kind of skills being demonstrated by students at particular locations on the underlying variable. Moreover, we assume the success rate odds of 50/50 at transition points are linked to a change in the required cognitive skill and thus hold implications for teaching strategies and resourcing.

The first point (item grouping) is justified on statistical and conceptual grounds if items have behaved in a Rasch-like manner. The second point (giving the set of skills a label) can be criticised, but only on conceptual, and not statistical, grounds. The conceptual criticism is only valid if the items within a group do not suggest a meaningful and unifying set of skills or competencies that are apparent to a professional observer. If there is not a single unifying and meaningful set of skills demanded by the set of items, the set may need to be adjusted to make the interpretation clearer. That is, some items may need to be omitted because, despite appropriate calibration and fit indices, the set of items may not be substantively relevant to the underlying construct or to cogent levels within the construct. Under these circumstances, they might not belong in the test at all and perhaps could be removed for the purposes of interpretation. This was not necessary in this case because the clusters did appear to offer an interpretable development sequence. It is a process we call *back translation* of the item set to the proposed construct and is consistent with Messick's (1989) view that validity is strongest when data fit closely to some underlying substantive theory

An important characteristic of the approach presented here is the reconstruction of the proposed construct underpinning the assessment. If the skills audit back translates to match or approximate the original proposed underpinning variable used to design and construct the assessment, it can also be used as evidence of construct validity. When this is linked to the index of item separation we have two pieces of evidence for construct validity (Griffin, 1990; Griffin & Nix 1991; Wright & Masters 1983). The technique has been used sparingly, but has emerged in several international studies. For example, Greaney and others used the procedure in their report on the Bangladesh testing in the "Education For All" project (Greaney, Khandker & Alam, 1995) in which they cited Griffin and Forwood's (1990) application of this strategy in adult literacy.

Knowledge Development Levels

Meaningful interpretations could be produced for seven knowledge levels based on common themes obtained from the item skills audit. Interpretation of levels identified the unfolding nature of students' increasing knowledge. Two criteria were used to identify and describe levels. First, there had to be identifiable sets of items and those sets needed to have a common substantive (conceptual) interpretation of underpinning skills. Grouping items on the variable map was a first step, but it was imprecise because of constraints of printers and line feeds, as some items placed on the same line have slightly different difficulties. The chart in Figure 7 illustrates where the difficulty of items changed.

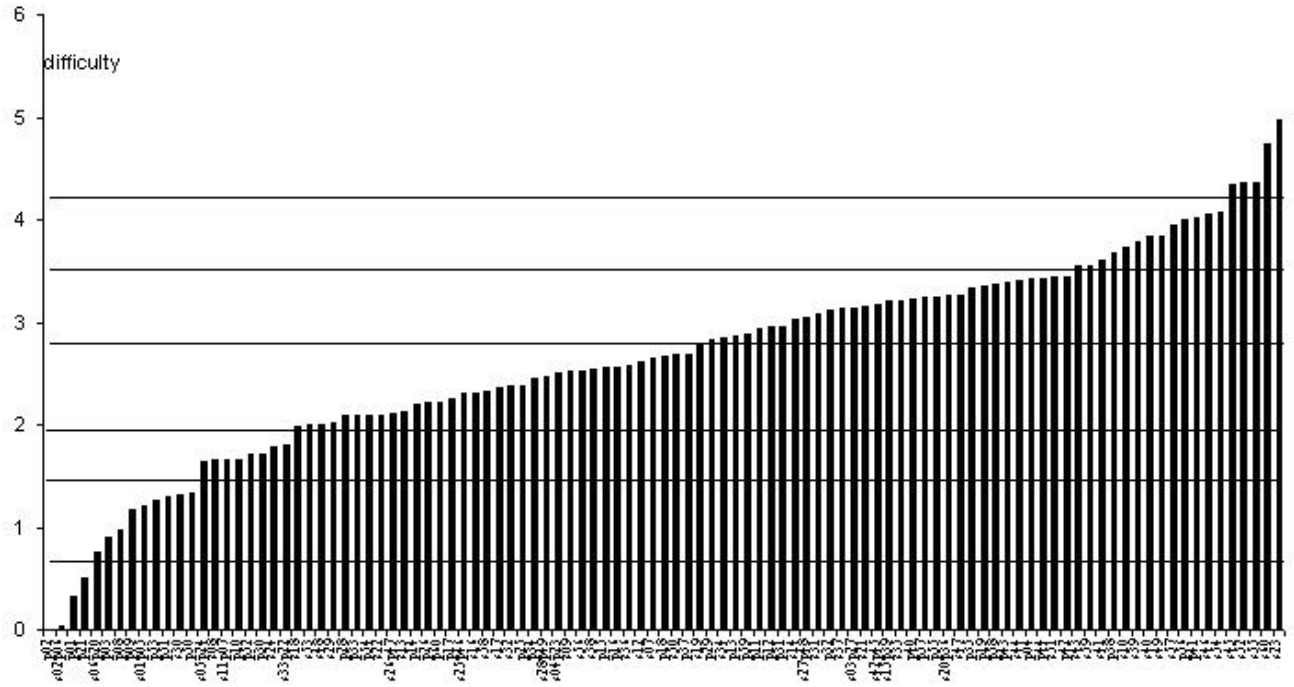

Figure 7. Knowledge test item order and level cut scores (Horizontal lines indicate cluster cut points).

Natural breaks in difficulty were identified and the items and cognitive descriptions were examined to determine whether a set with a common substantive interpretation could be found. A panel of specialists joined the project team for this exercise. Together they identified the breaks in the variable and then offered the substantive interpretation of the levels of competence. As a result of these approaches the cut levels for the shifts in skills were set at those shown in Table 7, and Table 8 shows the description of seven levels of students' knowledge and understanding. The mean and standard deviation of the logit scale were used to convert student scores to a common scale with a national mean of 500 and a standard deviation of 100.

Table 7
Cut Scores for Level Thresholds

Threshold	Knowledge and Understanding	
	Logit	K500
1 to 2	-2.15	214
2 to 3	-1.31	315
3 to 4	-0.67	391
4 to 5	+0.04	476
5 to 6	+0.79	566
6 to 7	+1.43	642

Table 8
Description of Levels of Knowledge and Understanding

Level	Description of the skill level
7	Highly developed knowledge base focusing on specialised understanding of local historical, cultural and contemporary issues.
6	Understands the impact of Asian historical figures on traditional and contemporary practices; Well developed knowledge of text styles, theatre, art and narratives associated with Asia.
5	Historical and contemporary events and their influence on Asia and Australia are understood, as are national and regional significance of festivals, celebrations and traditions in art, text language and theatre.
4	Specific knowledge of language, art, scripts, lifestyles, influential persons and stereotypes are contextualised within regional or national boundaries.
3	Links Asian icons to localised stereotypes in culture, people and religion Recognises diversity of Asia in terms of national industries and practices. Specific knowledge of food, land use, weather and regional industries emerges.
2	Recognises common icons associated with Asia and well known characteristics of language and customs
1	Recognises introductory ideas in symbols, food, customs, costumes and popular cultural artefacts and people independent of Asian cultures

Table 9 shows the knowledge (K500) scores that are linked to each knowledge level for Year 5 and Year 8 students, and Figure 8 shows the proportions of Year 5 and Year 8 students in knowledge levels.

Table 9
Levels of Year 5 and Year 8 students' knowledge and understanding about Asia

Level	1		2		3		4		5		6		7	
Year	5	8	5	8	5	8	5	8	5	8	5	8	5	8
Mean	181	182	283	288	363	357	438	439	516	522	597	598	683	694
SD	32	46	22	21	22	20	22	24	25	25	20	21	39	48

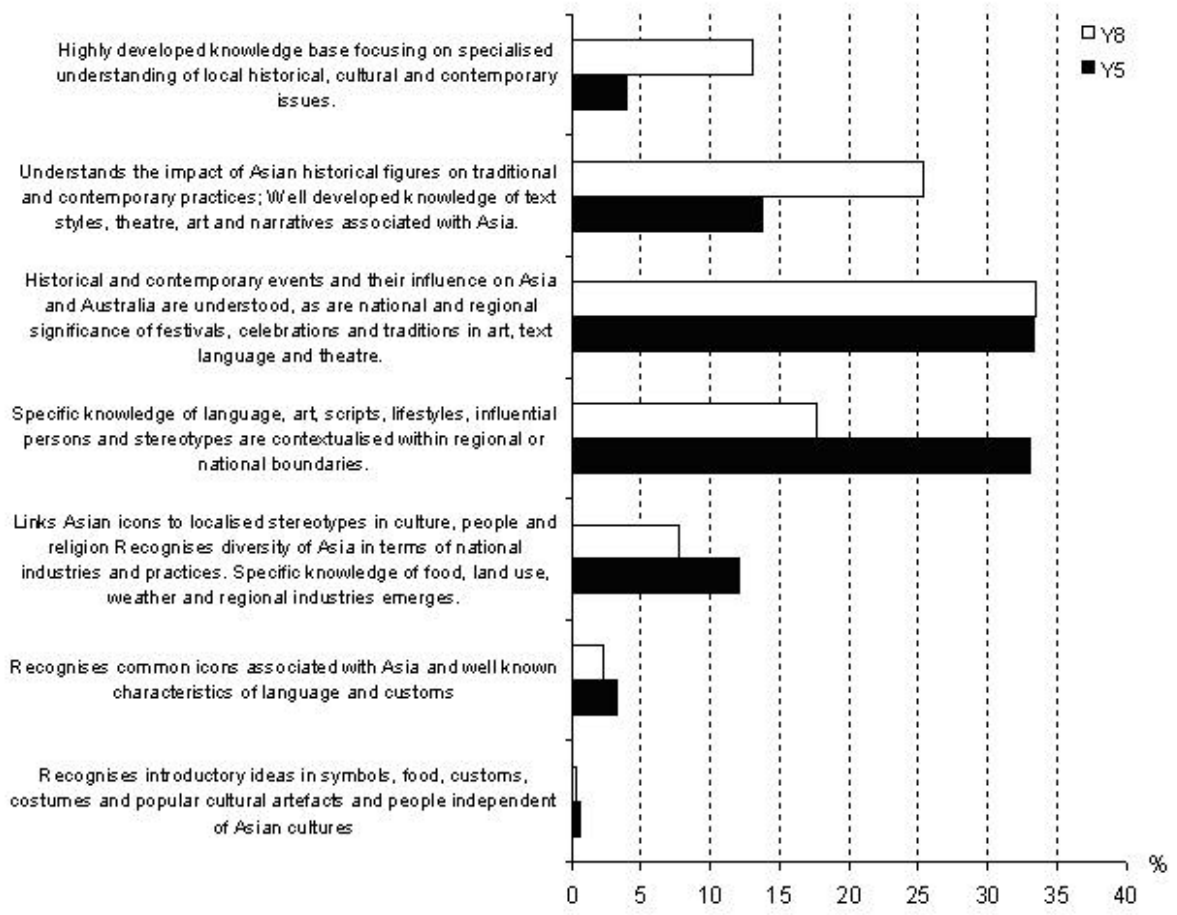

Figure 8. Distribution of students against ability levels.

The overall mean knowledge score for Year 5 students was 477, and 528 for Year 8 students. This means that the average Year 5 students' knowledge could be located at Level 4, and the average for Year 8 students was approximately Level 5.

Summary

This paper presents a national study that involved calibrating items distributed over 14 subtests, allowing for state and year level preferences to be selected from the common pool of 105 items representing three key learning areas. A single dominant dimension was identified using concurrent equating procedures and interpreted by a skills audit of items and an analysis of item groups on a metric of increasing difficulty. This procedure enabled the interpretation to move from a quantitative to a qualitative basis with direct and demonstrable implications for instruction. The impact of difference estimating procedures (pairwise and marginal maximum likelihood) was shown to be minimal in this instance when the proportion of misfitting persons and items was small.

References

Adams, R. J., & Khoo, S. K. (1995). *Quest: Interactive item analysis.* Melbourne: ACER.

Andrich, D., & Lau, G. (2002). Reconciling RUMM and Quest. Personal communication. July.

Baumgart, N., & Halse, C. (1999). *Asia Education Foundation: National evaluation of the Second Triennium.* Sydney: University of Western Sydney, Nepean School of Lifelong Learning and Educational Change.

Coates, H. (2001). Student non-response in cross-national studies of student achievement: Establishing psychometric, educational and political equivalence and equality. Paper presented at the Annual Faculty of Education Post Graduate Students' Conference. The University of Melbourne.

Greaney, V., Khandker, S. R., & Alam, M. (1990). *Bangladesh: Assessing basic learning skills.* Bangladesh Development Series. Dhaka: University Press. World Bank.

Griffin, P. (1990). Profiling literacy development: Monitoring the accumulation of reading skills. *Australian Journal of Education, 34,* 290-311.

Griffin, P. (1999). *Measuring achievement using a subtest from a common item pool: A cross national application of the Rasch model.* Melbourne: Assessment Research Centre.

Griffin, P., & Forwood, A. (1990). *Adult literacy and numeracy scales.* Canberra, Department of Employment Education and Training: Australian Government Printer.

Griffin, P., & Nix. P. (1991). *Assessment and reporting: A new approach.* Sydney: Harcourt Brace Jovanovich.

Griffin, P., Woods, K., Dulhunty, M., & Coates, H. (2002). *Australian students' knowledge and understanding of Asia.* Canberra: DEST.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13 – 103). New York: Macmillan.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Ross, K. (1995). *The Southern African Consortium for Monitoring Educational Quality.* Paris: UNESCO International Institute for Educational Planning.

RUMM Laboratory. (2001). *Rasch Unified Measurement Models.* Perth.

Wright, B., & Masters, G. (1983). *Rating scale analysis.* Chicago: MESA Press.

Zwinderman, A. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement, 19*(4), 369-375.